NPSOR-93-006

NAVAL POSTGRADUATE SCHOOL
Monterey, California

EMPIRICAL METHODS FOR ESTIMATING WORKLOAD CAPACITY

Michael P. Bailey

December 1992

Prepared for:
Naval Postgraduate School Research Foundation
Monterey, CA 93943

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral R. W. West, Jr.
Superintendent

Harrison Shull
Provost

This report was prepared for and funded by the Naval Postgraduate School Research Foundation, Monterey, California.

This report was prepared by:

MICHAEL P. BAILEY
Professor of Operations Research

Reviewed by:

PETER PURDUE
Professor and Chairman
Department of Operations Research

Released by:

PAUL J. MARTO
Dean of Research

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NPSOR-93-006	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Naval Postgraduate School 6b. OFFICE SYMBOL (*If applicable*) OR/BA	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (*City, State, and ZIP Code*) Monterey, CA 93943	7b. ADDRESS (*City, State, and ZIP Code*)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Naval Postgraduate School Research Foundation 8b. OFFICE SYMBOL (*If applicable*)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (*City, State, and ZIP Code*) Monterey, CA 93943	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO / WORK UNIT ACCESSION NO

11. TITLE (*Include Security Classification*)
Empirical Methods for Estimating Workload Capacity

12. PERSONAL AUTHOR(S)
Michael P. Bailey

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (*Year, month day*)	15. PAGE COUNT
Technical Report	FROM_____TO_____	December 1992	31

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

We explore experimental procedures for comparing the capabilities of complex discrete event service systems. Instead of measuring system capability by analyzing or simulating the system with a constant rate of arriving work, system capability is measured as the maximum rate of work arrival for which the system has a steady state. Hence, we seek the arrival rate which causes the system to be at full capacity. This rate is arguably the best indication of the service system's capability. We treat both work-conserving and non-work-conserving service systems, using traditional and specialized mesures of system performance.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT [x] UNCLASSIFIED/UNLIMITED [ ] SAME AS RPT. [ ] DTIC USERS	21. ABSTRACT SECURITY CLASSICIATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL M. P. Bailey	22b. TELEPHONE (*Include Area Code*) (408) 646-2085 22c. OFFICE SYMBOL OR/BA

Empirical Methods for Estimating Workload Capacity

Michael P. Bailey

Code OR/BA

Department of Operations Research Naval Postgraduate School

Monterey, CA 93943-5000

December 1, 1992

Abstract

We explore experimental procedures for comparing the capabilities of complex discrete event service systems. Instead of measuring system capability by analyzing or simulating the system with a constant rate of arriving work, system capability is measured as the maximum rate of work arrival for which the system has a steady state. Hence, we seek the arrival rate which causes the system to be at full capacity. This rate is arguably the best indication of the service system's capability. We treat both work-conserving and non-work-conserving service systems, using traditional and specialized measures of system performance.

Introduction

As industrial engineers, applied probabilists, simulationists, and systems analysts, we are often called upon to evaluate systems which service input traffic and produce finished products. These sytems are sometimes traditional queues or networks of queues, but are often systems with queue-like characteristics which cannot accurately be modeled as traditional queueing systems. In practice and in the literature, this evaluation is traditionally based on excercising a model of the service system

by subjecting it to a stream of input traffic and estimating or calculating some expected system performance measure.

We feel that this typical experimental design is lacking, and that the shortcomings stem from the arbitrary choice of the distribution of the input process. Especially problematic are cases where the service system being modeled does not currently exist, where worst-case behavior is sought, or where we wish to evaluate the system in situations which are not accessible for data collection. Practical service system analysis is interesting only when the service system is in an environment where the workload is high relative to the system's capability to serve. In all that follows, we are interested in finding the intensity of the input process that taxes the service system to the extremes of its capabilities, and using this intensity as a measure to compare systems.

1 The General Service System Model

The service systems considered all have the following features:

1. a centralized, controlable, nonlattice process which generates tasks at a rate λ per unit time;
2. tasks are admitted upon generation and processed by the system;
3. a completed task is ejected from the system;
4. the system has the capability to process as many as μ tasks per unit time.

We will call such systems Discrete Event Service Systems (DESSs), see figure 1. We will allow the system to create, combine, destroy, or absorb tasks – the DESS need not be work conserving. A DESS must, however, be mixed or open, as we are interested in how much work we can inject into the system without overburdening it. We may measure the performance of the system using traditional queueing measures such as the number of tasks resident in the system or some production cost, or we may opt to analyze some measures which are particular to the application at hand. In this work, we will approach the work-conserving and nonwork-conserving systems seperately.

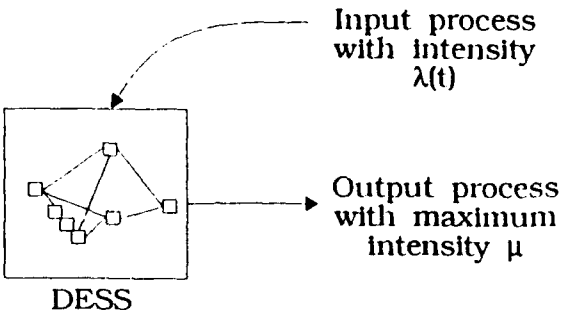

Figure 1: A simple DESS.

1.1 Motivating Example

Recently a model was constructed of all of the single-channel radio communications in a Marine Expeditionary Brigade (MEB). A MEB consists of approximately 20,000 marines who perform in three subelements, the Ground Combat Element (GCE), the Air Combat Element (ACE), and the Combat Service Support Element (CSSE). The single-channel radio network of the MEB may consist of several thousand radios operating on several hundred radio nets. The goal of the study was to allocate newly purchased antijamming radios to the different units in the MEB.

The radio traffic was modeled using the Marine Corps' version of a structured trafffic model. There are sixty classes of communication task packages, each representing a different mission that the MEB executes in battle. Each task package, called a Broad Operational Subtask (BOST), was composed of several tasks which were called Message Exchange Occurrances (MEOs). Each BOST contained between five (5) and sixteen (16) MEOs, and each MEO has a set of other MEOs in the BOST which must be completed before it can be initiated.

The MEOs are completed by transmittting a message from the specified sender to each of the specified receivers. If this communication cannot be accomplished using the specified doctrinal radio

net, the sender will attempt to reach the receiver using other nets, relays through other radios, or using messengers. The user may also elect to delay the MEO some short time before attempting it again. The radios in the system

- experience failures;
- become jammed by scenario-specified jammers;
- hold queues of MEOs with different priorities;
- are capable of moving into and out of range of other radios;
- can be used in voice or digital mode;
- exit and join different radio nets as they move around or attempt to reroute MEOs.

The antijamming radio fails less frequently and is harder to jam than the old radio. However, it is much more time consuming to enter a net with the new radio than with the old one. Finally, if an old radio tries to contact a new radio, the new radio must transit into a less capable mode in order to receive the MEO, then reenter his regular net of new radios.

Our model features structured traffic being presented to a set of servicing capabilities which act semi-autonomously. The routing of the MEOs, sometimes using several transmissions to accomplish a single MEO, makes this model extremely difficult to analyze. To expedite the analysis and to acheive maximum flexibility and sponsor acceptance, an object-oriented simulation model was constructed to allow analysts to build MEB radio net structures and test their capabilities against on another.

This brings us to the search for the right rate of generstion of BOSTs in the system. This generation rate is clearly dependent on the pace and nature of the battle being experienced by the MEB. After some initial searching through volumes of data from the recent Desert Storm operation, we found that the Marine Corps didn't take time during their ground war to record the time and type of every BOST they executed! Experience with the model showed that the measured performance depended greatly on the pace of the presented traffic. When considering various C^3I architectures,

we found the ranking of the radio allocations from best to worst changed as the BOST generation rate changed. The sponsor wanted to identify the best architecture for the most intense traffic.

2 Work Conserving Systems

We first study the behavior of systems where there is a one-to-one correspondence between the tasks we submit for processing and the finished tasks the service system produces. Work-conserving queueing models do not allow

- tasks to expire while in service;
- tasks to create other tasks while in service;
- tasks to be split or combined;
- tasks which never finish service.

Work-conserving queueing system models are common in both the practice and literature of applied probability. In a typical experiment, we generate input to the system at a constant rate, monitor the performance of the system either at fixed intervals or upon departure from the system, and employ well-known methods of steady-state analysis to estimate the steady-state average of the performance measure.

A maxim of the analysis of service systems is that the system will have stationary long-run behavior if and only if the number of arriving tasks are, on average, less than the number of tasks the system is capable of processing. If our overall system can work at a maximum of μ tasks per unit time, we can input as many as μ per unit time and the system will remain stationary. If λ is our arrival rate for the system, we wish to manipulate λ to expose μ.

2.1 Generating Data

There are two ways we can generate data from a work-conserving system which will reveal the maximum processing rate in the system. They are:

- input tasks to the system at a rate known to be much higher than the system can handle;
- fill the system, then input a new task every time that a task completes.

Instead of choosing a very high input rate and dealing with the problems of exploding buffer contents and a nonrecurrent system, we will simply close off the system and recirculate the tasks which finish. This approach will also serve as a good introduction to the analysis of systems which do not conserve work.

Thus, we examine a special kind of closed queueing network - one with a single loop-back which all tasks traverse. Let $\lambda(t)$ be the time-dependent rate of recirculation of tasks in the system. So long as the system contains enough tasks to keep it working at capacity, we have $\lambda(t) \rightarrow \mu$ as $t \rightarrow \infty$. Kelly [3] and Walrand [9] both show this for exponentially distributed service, and Disney and Kiessler [2] make the extension to Jackson networks. The result can be extended in the obvious way by treating Phase-type distributions for service times, to produce the result we seek ($\lambda(t) \rightarrow \mu$ as $t \rightarrow \infty$) for generally distributed service.

Example

We will demonstrate this method on the Jackson network shown in figure 2.

2.2 Detecting Transition to Steady State

Let us simulate the completion of the first N customers serviced by the closed system for M independent replications. Let $T_{i,j}$ be the j^{th} time between recirculation during the i^{th} replication. Thus, $T_{i,j}, i = 1, 2, \ldots, M$ is a set of *iid* samples. Let $\bar{T}_j = \sum_{i=1}^{M} T_{i,j}/M$ be the average recirculation time process. We seek the index N^* such that $ET_{i,j} = E\bar{T}_j = \mu$ for all $j > N*$. Hence, we are in the setting of a traditional initial transient detection problem.

There exist many ways to tackle this problem, including

- cross-replication confidence intervals, [10]
- tests for significant drift;

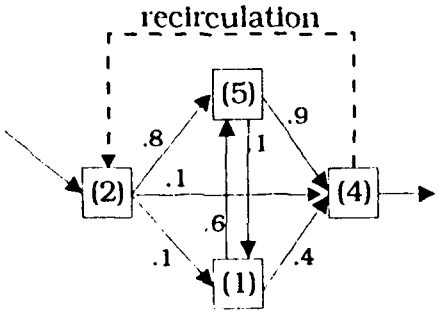

Figure 2: A Jackson Network designed to have a maximum service rate of 0.5. The numbers in parentheses are the number of servers at each station, and the routing probabilities are shown on the workstation connections. All servers have unit service time, and all buffers are infinite. The dashed line shows the recirculation route added to force the system to serve at the maximum rate.

- standardized time series (STS), [6].

In our experiences we have found STS useful, especially in a slightly modified version we have developed, which we call *ratio STS* (RSTS).

2.3 Ratio STS

The method of standardized time series (STS) [6], produces confidence intervals from autocorrelated, stationary data. This method was used in [7] to detect the existence of initialization bias in simulation output, and was sharpened to produce optimal tests when the functional form of the initialization bias is known.

Suppose that we have M independent samples of n points each, with $Y_{i,j}$ being the j^{th} point in the i^{th} independent sample. Let

$$\bar{Y}_{i,j} = j^{-1} \sum_{k=1}^{j} Y_{i,k} \tag{1}$$

for $i = 1, 2, \ldots, M$, and with $\bar{Y}_{i,0} = 0$ for each i. The time series $S_i(k), k = 1, 2, \ldots, n$ is constructed for each independent replication i as

$$S_i(k) = \begin{cases} \bar{Y}_{i,n} - \bar{Y}_{i,k} & for\ 0 < k < n \\ 0 & k = 0, n. \end{cases} \tag{2}$$

Let σ be the variance of Y_i. If $S_i(k)$ is divided by $\sigma\sqrt{n}/k$ and scale the index k so that the result resides in the unit interval $[0, 1]$, the resulting time series $T_i(t), 0 \leq t \leq 1$ is known to approximate a Brownian bridge as $n \rightarrow \infty$. This is the fundamental result of [6], and the theoretical basis of this sequencial procedure.

Schruben shows that scaling and summing $T_i(t)$,

$$A_i = \sigma\sqrt{n} \sum_{k=1}^{n} T_i(kn) \tag{3}$$

results in a normal random variable A_i with variance given by

$$VAR(A_i) = \frac{\sigma^2 n(n^2 - 1)}{12}. \tag{4}$$

Note that, except for a factor of σ^2, $VAR(A_i)$ is independent of the data, it relies only on the parameters of the experiment. Hence, for any integer $d < M$,

$$\chi_d^2 \sim \sum_{i=1}^{d} \left(\frac{A_i}{\sqrt{VAR(A_i)}}\right)^2 \tag{5}$$

$$= \frac{12}{\sigma^2 n(n^2-1)} \sum_{i=1}^{d} A_i^2. \tag{6}$$

The original STS used to detect initial transients used χ_d^2 as a test statistic for stationarity of the mean response. If we form a ratio of χ_d^2 and χ_{M-d}^2, we can eliminate the need to estimate σ^2, forming

$$F_{d,M-d} \sim \frac{d^{-1}\sum_{i=1}^{d} A_i^2}{(M-d)^{-1}\sum_{i=M-d}^{M} A_i^2}. \tag{7}$$

This test statistic, which we call the RSTS test statistic, is easy to use in all of the applications where STS is applied. In particular, if we are interested in determining the onset of steady state, we can form the backward-moving sequences $A_{i,j}, j = n-1, n-2, \ldots, 1$ for each replication i, where $A_{i,j}$ is formed from the subsequence $Y_{i,k}, k = j, j+1, \ldots, n$, the portion of the i^{th} replication between j and n. Thus, we form the sequence of F-statistics

$$F_{d,M-d}(j) \sim \frac{d^{-1}\sum_{i=1}^{d} A_{i,j}^2}{(M-d)^{-1}\sum_{i=M-d}^{M} A_{i,j}^2}. \tag{8}$$

If we assume that the system is in steady state when each of the $A_{i,n}$ are collected, then we can detect the transition of the system into steady state by looking at the first index N^* where $F_{d,M-d}(N*)$ exceeds the critical value for an F random variable with identical degrees of freedom. This method is demonstrated in the following example.

2.3.1 Example

Continuing with the work-conserving system example, suppose that we

- start the system with 25 tasks enqued at workstation 1 at time 0.0;
- simulate $N = 500$ customer recirculations;
- replicate $M = 20$ times.

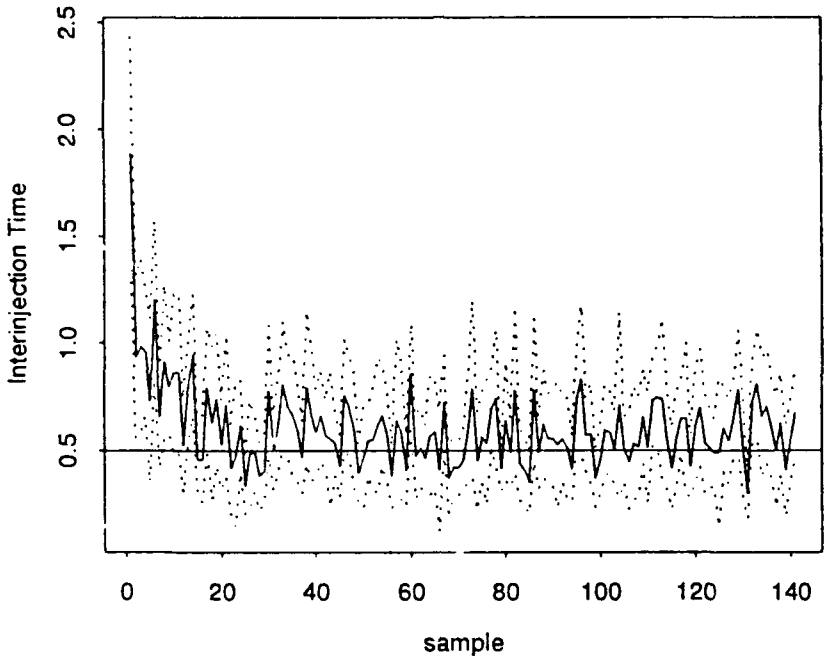

Figure 3: Trajectory of the mean process $\bar{T}_j$ and the confidence intervals.

Figure 3 shows the trajectory of $\bar{T}_j$ and the associated confidence interval process for the first 140 recirculations. Clearly, by sample $N* = 80$ we have passed the criteria for being in steady state according to Welsh's cross-replication confidence interval method. Furthermore, we can see that any detectable slope in the mean process is neglegible. When tested for our 20 independent samples, the drift of tested to be insignificant (H_0: no drift has p-value ≈ 0.4).

The mean time between recirculations in $100, 101, \ldots, 140$ was $\bar{T} = 0.56$, (confidence interval $(0.55330, 0.56691)$), clearly not as fast as the $\mu = 0.50$ which we know to be the system's capacity. Performing unweighted RSTS in the first 140 samples showed *no transition to steady state detectable* – the procedure seemed to be accurate enough to discern that the transition had not yet occured, see figure 4.

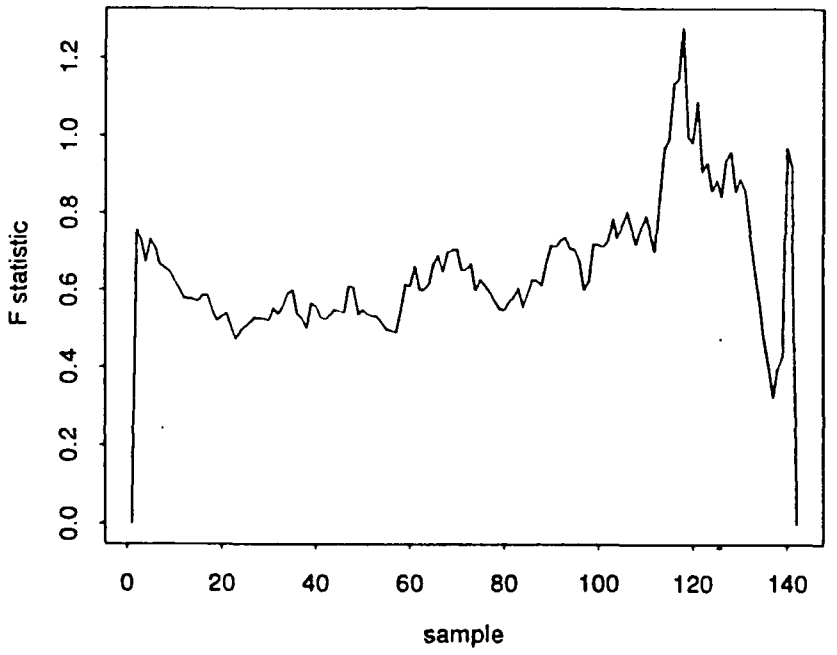

Figure 4: RSTS performed on the first 140 sample recirculation times, using 12 numerator degrees of freedom and 8 denominator degrees of freedom. Conclusion: no transition to steady state was evident.

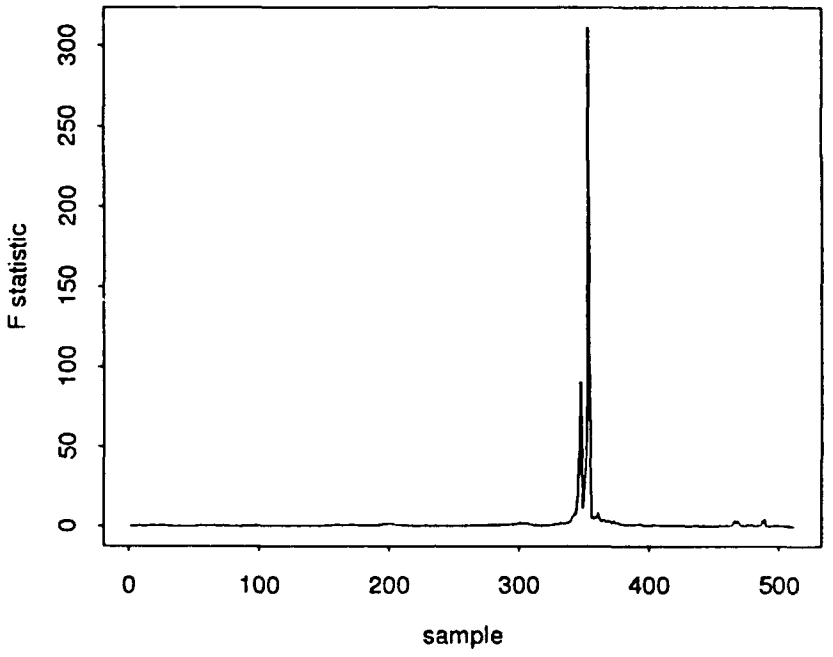

Figure 5: RSTS performed on the first 500 sample recirculation times.

When we extend the length of the runs we consider to the full 500 samples, we see that RSTS was able to indicate a strong transition to steady state around the $N* = 330$ sample, see figure 5. Averaging the samples collected in $350, 351, \ldots, 500$, we observe an overall average of $\bar{T} = 0.501$ (confidence interval $(0.49816, 0.50389)$).

RSTS clearly dominated the other traditional initial transient methods. In the case of the recirculating jobs, we clearly have a very gradual descent to the steady-state average. The detection method is not important to our overall theme, though we must issue a general caution: The choice of N^* should be made very conservatively.

3 Non-Work-Conserving Systems

In this section, we consider the case where work is not conserved in the system, and where the measure of performance is very general. This case, which is much more interesting and applicable than the work-conserving case, allows us to treat cases where the service system may share unusual characteristics with the real system, and where the measure of performance of the system may be dictated by the study sponsor.

3.1 Motivating Example, Revisited

In the motivating example, BOSTs were input to the system. Depending on the BOST type, the BOST might

- splinter into several communications tasks, which may splinter further at a later time;
- require partial or full reassembly at different points;
- expire after is has reached a certain age.

Thus, the system clearly doesn't conserve work.

The sponsor was interested in specifying

- a mix of different BOST types which the input was comprised of;
- a time allotment for each type of BOST, and a time when the BOST expires and is removed from the system;
- a one-time cost, by BOST type, assessed when the time allotment expires with the BOST still in process;

All of these requirements were imposed because of the need to assess the system's ability to handle communications as diverse as artillery targetting and mission execution traffic, medical evacuation requests, situation reports, intelligence traffic, logistics and administrative communications, and communications allowing the radio nets to counter radio jamming.

For this application, we constructed a penalty process $p(t)$, which superimposed lateness and expiration penalties from all of the traffic in the system, and accumulated these penalties in $[0,t]$. We compared C^{3}I architectures based on the rate-of-climb of this penalty when traffic was inserted into the system. The sponsor really wanted to know the answer to the following question: "Which architecture has the best peak-load performance?" Since the model was an idealization of the real system, and since wartime traffic load data is not available, we were faced with the problem of determining what traffic rate represented peak loading to the system.

3.1.1 The Workload Ramp

Given that we do not know the service rate of the system, we can try to produce an estimate by modulating the intensity of the system's input and observing the effect this has on the system performance. One might attempt this by stepping through some reasonable intensity values, or by doing some sort of iterative search. After considering several alternatives, we decided that a nonhomogenius input process with gradually increasing intensity might be appropriate. This idea is similar to testing a stereo system for its ability to play loud music - we gradually turn up the volume, listening for the point where the music begins to become distorted.

The mechanics of generating the ramping workload process and calculating the likelihood ratio for a generated sample path are now presented. The requirement is to generate a nonhomogeneous Poisson process with jumppoints $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N$ from an intensity function $\lambda(t)$ given by

$$\lambda(t) = \begin{cases} \lambda(0) + rt & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{9}$$

where $\lambda(0) > 0$ is the initial intensity and r is the rate of climb of the intensity function. In this presentation $sign(r)$ is not specified, and is a point of interest in future research. Let $N(t)$ be the number of tasks arriving during $[0,t]$. From the above equation, we have

$$a(t) = E[N(t)] \tag{10}$$

$$= \int_0^t [\lambda(0) + rs]ds \tag{11}$$

$$= \lambda(0)t + rt^2/2 \tag{12}$$

Thus, if $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N$ is generated via (9), then $a(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N)$ is a Poisson process with rate 1. [1]. The scheme for generating our ramping workload process is given as

$\tau_0 = 0.0, j = 1$

while NOT DONE

generate U $\sim$ **U[0, 1]**

$\tau_j \leftarrow \tau_{j-1}$ - **ln(U)**

$\tilde{x}_j \leftarrow a^{-1}(\tau_j) = \frac{-\lambda(0)+\sqrt{\lambda(0)^2+2\tau_j r}}{r}$

end while

Algorithm 1: The generation of the ramping intensity workload process.

Let $\bar{G}(x) = 1 - G(x)$ be the complement of the joint distribution of $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_i, \ldots$ Then

$$\bar{G}_{\tilde{x}_i|\tilde{x}_{i-1}}(t) = P[\tilde{x}_i - \tilde{x}_{i-1} > t|\tilde{x}_{i-1}] \tag{13}$$

$$= e^{-[a(\tilde{x}_{i-1}+t)-a(\tilde{x}_{i-1})]} \tag{14}$$

$$= e^{-[\lambda(0)(\tilde{x}_{i-1}+t)+r/2(\tilde{x}_{i-1}+t)^2]-[\lambda(0)\tilde{x}_{i-1}+r/2(\tilde{x}_{i-1})^2]}. \tag{15}$$

Thus, the conditional density function $g_{\tilde{x}_i|\tilde{x}_{i-1}}(t)$ is given by

$$g_{\tilde{x}_i|\tilde{x}_{i-1}}(t) = [\lambda(0) + r(t + \tilde{x}_{i-1})]e^{-[\lambda(0)t+r/2(t^2+2\tilde{x}_{i-1}t)]} \tag{16}$$

$$= [\lambda(0) + r(t + \tilde{x}_{i-1})]e^{-[\lambda(0)+r(t/2+\tilde{x}_{i-1})]t}. \tag{17}$$

This last presentation of $g_{\tilde{x}_i|\tilde{x}_{i-1}}(t)$ highlights the nature of the density function, with leading constant given as $\lambda(0) + r\tilde{x}_i$, the rate at the time of the generation epoch, and the exponent given as $-t$ times $\lambda(0) + r(\tilde{x}_{i-1} + \tilde{x}_i)/2$.

When appropriately calibrated, the ramping workload process will drive the DESS into regimes in which it is underutilized, progressing to the point of total utilization, and then becoming saturated.

3.2 Detecting the Transition to Overloaded for a DESS

Let our system performance measure for input intensity λ and time t have expected value $\psi(\lambda, t)$. Our only assumptions about ψ are that it is responsive to changes in λ, it grows at a rate similar to a degree s polynomial when $\lambda < \mu$, and it grows faster when $\lambda > \mu$.

Hence, by estimating the $s+1^{st}$ derivative of the performance measure ($s+1$ because we would like to deal with a mean-zero sequence), we can produce a sequence with

- constant mean when $\lambda < \mu$,
- some drift when $\lambda > \mu$.

Let $t_1, t_2, \ldots, t_n$ be n evenly spaced points in time, where $\lambda(t_1)$ is believed to be less than the service capacity μ of the DESS, and $\lambda(t_n)$ is believed to be much greater than μ. Our method for estimating DESS capacity will

1. replicate the system performance M times using the workload ramp as the input process, collecting data $\Psi_i(\lambda(t_j), t_j)$ for the i^{th} replication;
2. form the $s+1^{st}$ derivative data $\Psi_i^{(s+1)}(\lambda(t_j), t_j), j = s+1, s+2, \ldots, n, i = 1, 2, \ldots, n$ using sequencial differences;
3. perform a transition detection to determine the point N^* where $\Psi_i^{(s+1)}(\lambda(t_j), t_j)$ no longer have constant mean 0.

We will propose and evaluate three methods for detecting the transition in the performance measure:

- modified $\bar{X}$-charts;
- RSTS;
- adaptive regression splines.

The application of each of these three methods will be made more difficult by a common feature of data we have collected – although the mean performance becomes constant after several differencing

WORKSTATION	P[DESTRUCTION]	P[CREATION]	INSERTION STATION
left	0.2	0.0	-
top	0.1	0.4	2
bottom	0.2	0.1	4
right	-	0.3	2

Table 1: Destruction and Creation of Tasks within the DESS. Tasks that are destroyed are not considered completed.

operations, the variance still grows. Hence, our model of the data from the system when unsaturated will be $Y_{i,j}, i = 1, 2, \ldots, M, j = 1, 2, \ldots, n$, which are mutually independent, and where $EY = 0$ is constant and σ_Y is unknown and assumed to *vary*.

Example, Continued

We modified the Jackson network described in section 2 so that it was no longer work-conserving, and so that it exhibited properties similar to the communications system described above. Table 1 shows what can happen to tasks in the system when they complete service at each of the nodes.

For this system, we still have a maximum input rate of $\lambda = 2$, as workstation 1 is not interfered with, and still has two servers serving at unit speed. The tasks for the system are of three classes. All have identical processing speeds at the workstations, but each class pays a different price for waiting in each workstation buffer. Finally each task can stay in the DESS cost-free for some period of time. After this delay, the cost per unit time is accumulated based on the buffer the task resides in. Each task becomes costless once it leaves the system. All of the data on task classes is in table 2.

The system's measure of performance is the cost accumulated from the beginning of the simulation. We subjected this system to a ramped workload process which started with insertion rate $\lambda(0) = 1.0$ and climbing at a rate $r = 0.00666$. Thus, the system capacity is reached at time

CLASS	FREE TIME	COST(LEFT)	COST(TOP)	COST(BOTTOM)	COST(RIGHT)
1	2.0	4	1	0	1
2	0.0	1	2	3	4
3	1.0	3	2	1	0

Table 2: Free Time and Holding Costs for Task Classes.

$150 = N^*$. We continued each experiment to time 300.

As we see from figure 6, we have the expected properties of constant *mean* and *growing variance* for the second derivative of the accumulated costs when $t < 150$, and something else occurring when $t > 150$.

3.2.1 Control Charting

The most straightforward methodology comes from the field of quality assurance, [5], and involves the use of a sequencial hypothesis test. We wish to know the first time that we can conclude that it is no longer plausible that the response mean is $EY = 0$. Our growing varaince causes us to take one of two actions:

- use the cross-replication sample standard deviation to form a confidence interval;
- model the growth of the standard deviation and use the model when setting control limits.

Using the second derivative data from our example system, we can see that the former method is inclusive because of constant false alarms (the lower control limit makes several visits above the x-axis during the trajectory), while using the modeled standard deviation creates an envelope which the mean stays inside even when we know the system is exhibiting drift, see figure 8.

3.2.2 RSTS for Detection of Saturation

Let us return to the construction of the sequencial RSTS methods. In this case, we have two alterations to make:

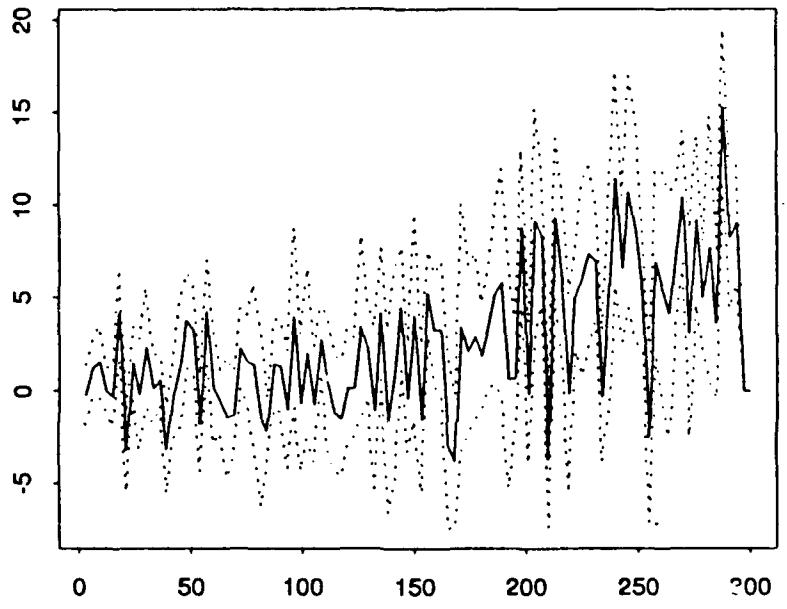

Figure 6: Trajectories of the Second Derivative of the Accumulated Cost Samples. The center line is the mean of $M = 20$ independent replications, while the surrounding lines are the upper and lower normal confidence intervals.

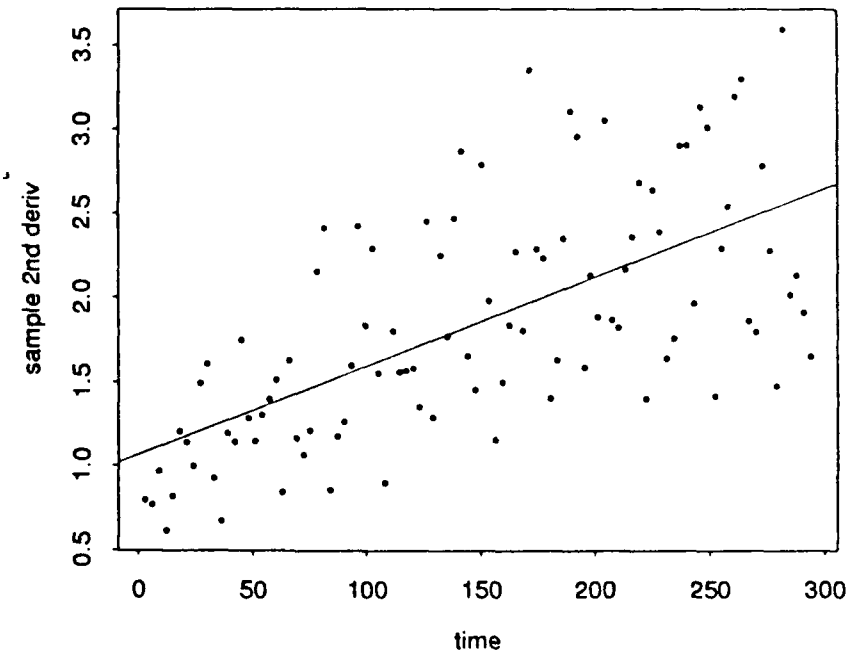

Figure 7: The sample standard deviation of the responses and a regression model of the growing standard deviation.

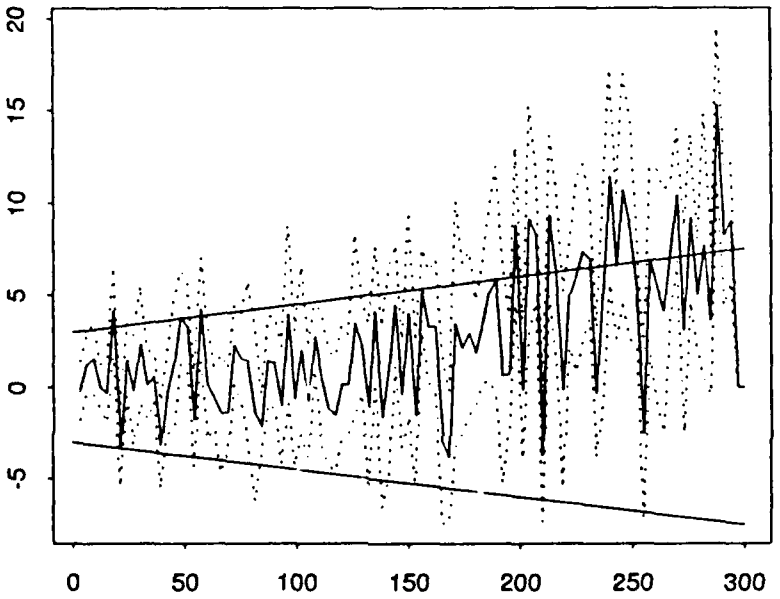

Figure 8: Control limit detection methods.

1. the process must be reversed to detect transitions out of steady state;
2. the growing variability of the data violate the assumptions under which $T_i(t)$ on $[0,1]$ was shown to be a Brownian bridge - the underpinning of the method.

Let the sequences $A_{i,j}, j = 1, 2, \ldots, j$ for each replication i, where $A_{i,j}$ is formed from the subsequence $Y_{i,k}, k = 1, 2, \ldots, j$. Thus, we are moving through the output sequence in the forward direction (opposite the usual STS for initialization bias).

The growing variability of the output must be attacked directly. The model which fits the output with $\lambda < \mu$ is a mean-zero model with linearly increasing standard deviation. The expected number of input tasks, $a(t)$, is also quadratically increasing and is intimately linked to the growth of the preformance measure and its variability. If we collect samples of the cost function $c(\lambda, t)$ such that there are a constant expected number of input tasks per sample, we can avoid the problem of growing variability. Our data will still be formed by taking sequencial differences, but the intervals of sampling will contract. Figure 9 shows empirically that sampling this way produces data with constant standard deviation in our example. Proving that this technique works in all cases is impossible because of the breadth of our generality here.

Let us establish the time interval sequence $t_1, t_2, \ldots, t_n$ such that we expect to inject exactly C tasks into the system between t_i and $t_{i+1}, i = 1, 2, \ldots, n-1$. Hence, given t_i we can compute $t_{i+1} = t_i + \Delta t$ using

$$a(t_i + \Delta t) - a(t_i) = C. \tag{18}$$

Using 12, we produce the equation

$$\lambda(0)\Delta t + \frac{r}{2}(2t_i\Delta t + \Delta t^2) - C = 0, \tag{19}$$

leading to

$$\Delta t = \frac{-(rt_i + \lambda(0)) + \sqrt{r^2t_i^2 + 2rt_i\lambda(0) + \lambda(0)^2 + 2rC}}{r}. \tag{20}$$

If we sample the DESS cost function, starting with some initial point t_1 and with some arbitrary value C, we produce a sample with constant standard deviation. Taking the sequencial differences as above, we produce a sequence which has the properties required to perform RSTS.

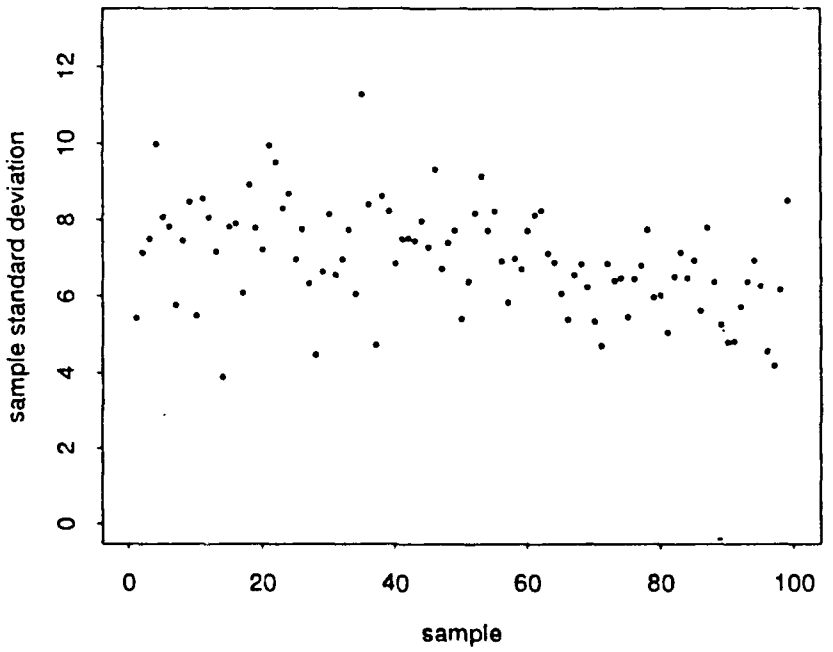

Figure 9: Sample standard deviations from data collected using the modified sampling points method.

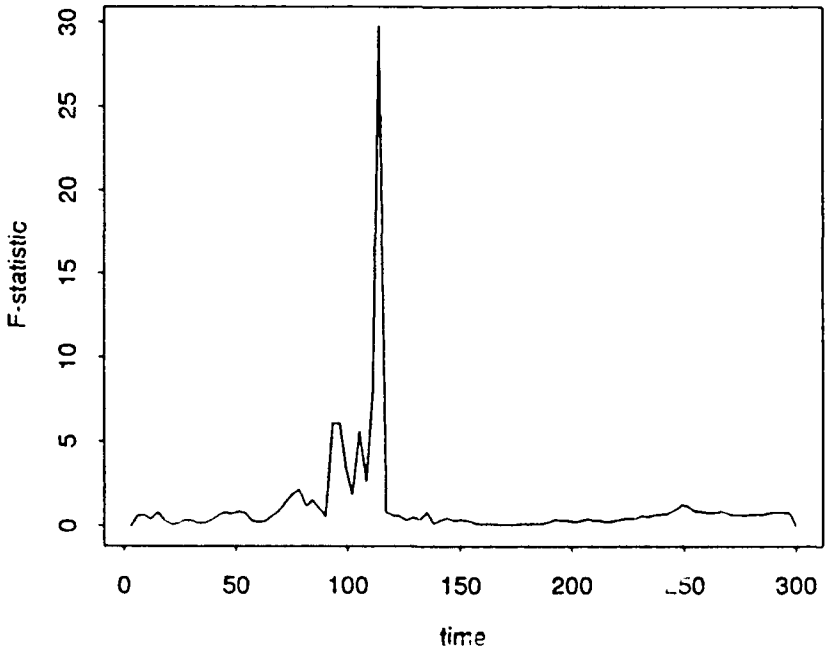

Figure 10: Trajectory for RSTS for the modified sampling points data collection. Result: clear transition at time 123.0

In our example system, we used the modified sampling points to produce the F-statistic trajectory shown in figure 10. The RSTS method gave a clear signal that the system lost steady-state at time 123.0, where $\lambda(t) = 1.82$ and $\rho = 0.910$. We also tried RSTS with evenly spaced points, and achieved approximately the same result. From this experience, we conclude that straightforward RSTS is fairly robust with respect to fluctuations or growth in variability when the changes are coordinated as they are in this experiment. Analytical investigation has shown that these fluctuations *do not cancel* in the RSTS test statistic.

3.2.3 Adaptive Regression Splines

Adaptive regression splines, especially multivariate regression splines, are the focus of intense basic research, see [8]. In our application, the regression spline required is especially simple, as the model is of a single independent variable, and we seek a single knot in the regression spline at the point where the zero-mean model departs from the data. Using the methods in Larson [4], we can derive the location of this single knot analytically.

The model used is stated as

$$Y(t) = \begin{cases} \beta_0 + \beta_1(t - N^*) + \epsilon_t & t \leq N^* \\ \beta_0 + \beta_2(t - N^*) + \epsilon_t & t > N^*. \end{cases} \tag{21}$$

where N^* is still our point where the regression model changes. If we further prescribe that

$$\beta_0 = 0, \tag{22}$$

$$\beta_1 = 0. \tag{23}$$

we will be fitting the model with zero mean to the left of the single knot. Let

$$SSE_L = \sum_{t \leq N^*} Y^2(t); \tag{24}$$

$$SSE_R = \sum_{t > N^*} (Y(t) - \hat{\beta}_2(t - N^*))^2; \tag{25}$$

$$SSE = SSE_L + SSE_R. \tag{26}$$

The method involves the optimal location of N^* to minimize SSE. The procedure provided in [4] can be simplified for our application into a single-pass examination of the means of the data, but is valid only when a linear model is appropriate for the data to the right of N^*. This method is valid when there is growing variability, as seen in our application.

In our example, we fit the model in (21) with three parameters, then restricted it to the mean-zero model prescribed in (22-23). Figure 11 shows the two regression models. The three parameter model located a knot at time $N^* = 136.25$, when the input rate is 1.91 and the traffic intensity is 0.95. With the one parameter restriction, the adaptive spline located the optimal knot at 128.0. We

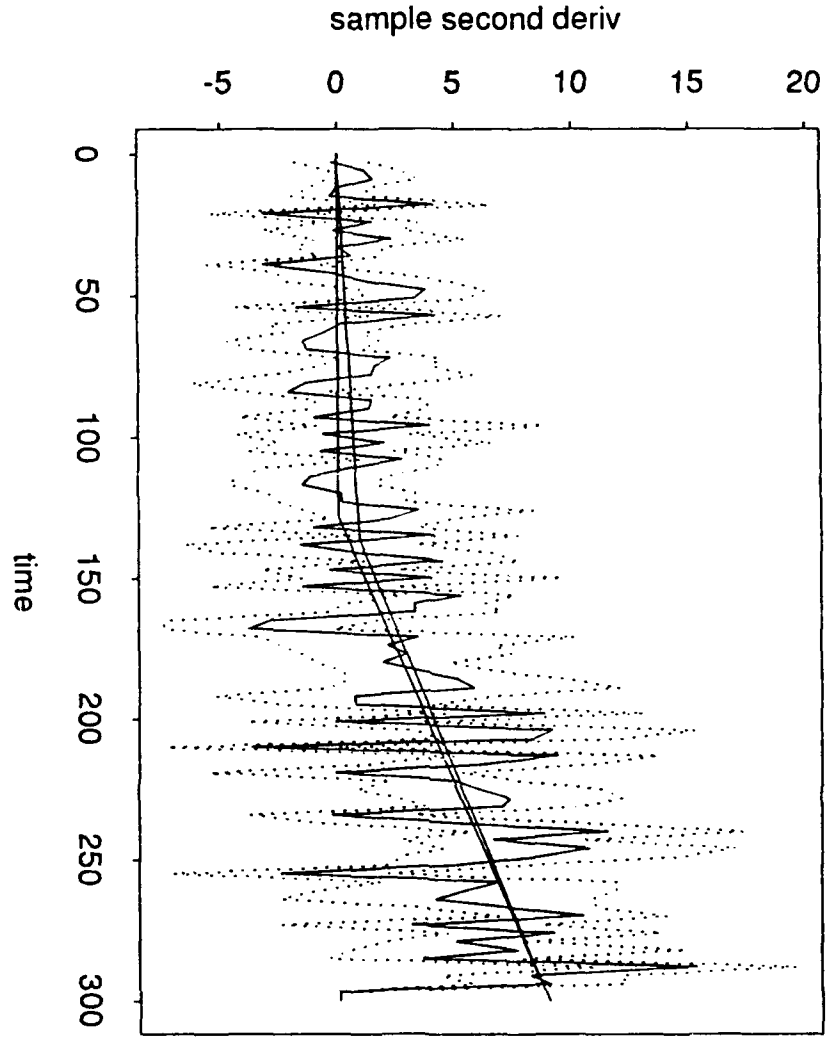

Figure 11: Adaptive regression spline with a single knot, three parameter and one parameter models.

should note that the linear model does fit fairly well to the right of N^*, as one might expect of any queuing application with less capacity than required.

4 Conclusion

In this work, we have described the various problems which arise when we are attempting to determine the service capacity of a *black box* service system. We divided this investigation into two distinct parts, one dealing with simple queueing systems which conserve work. In this case, we

showed the advantages of closing the system so that output from the system was recirculated, as the *time between recirculations converges* to the system's service rate. We investigated ways to detect this convergence, and showed how difficult this is in a simple Jackson network example.

The second part of the exploration dealt with systems which

- do not conserve work;
- have their performance measured using a general holding cost mechanism or some other performance measure.

In this case, we showed that if we modulated the input process using a ramped-intensity workload process, we could drive the system from underutilized to saturated. We explored three methods which unveil the point where this transition takes place, and demonstrated each on an example.

The wider significance of this work is the beginning of an exploration for empirical methods *for determining the capacity of a service system*. This exploration is done not by representing the system using a queuing model which we know how to analyze *a priori*, but by using a realistic model of the system and measuring its performance in terms the *user* has in mind.

Acknowledgements

This work has been supported by the Naval Postgraduate School Research Foundation, and benefited from previous support from the U. S. Marine Corps. The author would like to acknowledge the generous contributions of Professors William Kemple and Harold Larson of the Naval Postgraduate School, and the encouragement and insightful comments of Professor Lee Schruben of Cornell University.

References

[1] Bratley, Paul, Bennett Fox, and Linus Schrage. 1983. *A Guide to Simulation*, Springer-Verlag, New York.

[2] Disney, R. L. and P. C. Keissler. 1987. *Traffic Processes in Queuing Networks, A Markov Renewal Approach.* Baltimore, Maryland: Johns Hopkins University Press.

[3] Kelly, F. P. 1979. *Reversibility and Stochastic Networks.* New York: John Wiley and Sons.

[4] Larson, H. J. 1992. Least squares estimation of linear splines with unknown knot locations. *Computational Statistics and Data Analysis.* **13**, p. 1-8.

[5] Montgomery, D. R. 1985. *Introduction to Statistical Quality Control.* New York: John Wiley and Sons.

[6] Schruben, L. 1982. Detecting initialization bias in simulation experiments. *Operations Research* **30**, p. 569-590.

[7] Schruben, L., H. Singh, and L. Tierney. 1983. Optimal tests for the initialization bias in simulation output. *Operations Research* **31**, p. 1167-78.

[8] Tong, H. 1990. *Nonlinear Time Series.* London: Oxford University Press.

[9] Walrand, J. 1988. *An Introduction to Queueing Networks.* Englewood Cliffs, New Jersey: Prentice Hall.

[10] Welsh, P. 1983. The statistical analysis of simulation results. *Computer Performance Modeling Handbook.* New York: Academic Press.

INITIAL DISTRIBUTION LIST

1. Library(Code 52) Naval Postgraduate School Monterey, CA 93943-5000	2
2. Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
3. Office of Research Administration (Code 08) Naval POstgraduate School Monterey, CA 93943-5000	1
4. Department of Operations Research Code OR Naval Postgraduate School Monterey, CA 93943-5000	1
5. Prof. Peter Purdue Code OR/PD Naval Postgraduate School Monterey, CA 93943-5000	1
6. Professor William Kemple Code OR/KE Naval Postgraduate School Monterey, CA 93943-5000	1
7. Professor Harold Larson Code OR/LA Naval Postgraduate School Monterey, CA 93943-5000	1
8. Professor Peter Lewis Code OR/LE Naval Postgraduate School Monterey, CA 93943-5000	1
9. Dr. Alfred George Brandstein, WF-SA MCCDC Quantico, VA 22134-5001	1
10. Center for Naval Analyses 4401 Ford Avenue Alexandria, VA 22302-0268	1